

Learning Super-Resolution Jointly from External and Internal Examples

Zhangyang Wang[†], Yingzhen Yang[†], Zhaowen Wang[‡], Shiyu Chang[†], Jianchao Yang[‡], Thomas S. Huang[†]

[†]Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[‡]Adobe Systems Inc, San Jose, CA 95110, USA

{zwang119, yyang58}@illinois.edu, zhawang@adobe.com, chang87@illinois.edu, jiayang@adobe.com,
t-huang1@illinois.edu

Abstract—Single image super-resolution (SR) aims to estimate a high-resolution (HR) image from a low-resolution (LR) input. Image priors are commonly learned to regularize the otherwise seriously ill-posed SR problem, either using external LR-HR pairs or internal similar patterns. We propose joint SR to adaptively combine the advantages of both external and internal SR methods. We define two loss functions using sparse coding based external examples, and epitomic matching based on internal examples, as well as a corresponding adaptive weight to automatically balance their contributions according to their reconstruction errors. Extensive SR results demonstrate the effectiveness of the proposed method over the existing state-of-the-art methods, and is also verified by our subjective evaluation study.

Index Terms—Super-resolution, example-based methods, sparse coding, epitome

EDICS Category: TEC-ISR Interpolation, Super-Resolution, and Mosaicing

I. INTRODUCTION

Super-resolution (SR) algorithms aim to constructing a high-resolution (HR) image from one or multiple low-resolution (LR) input frames [1]. This problem is essentially ill-posed because much information is lost in the HR to LR degradation process. Thus SR has to refer to strong image priors, that range from the simplest analytical smoothness assumptions, to more sophisticated statistical and structural priors learned from natural images [2], [3], [4], [5]. The most popular single image SR methods rely on example-based learning techniques. Classical example-based methods learn the mapping between LR and HR image patches, from a large and representative external set of image pairs, and is thus denoted as *external SR*. Meanwhile, images generally possess a great amount of self-similarities; such a self-similarity property motivates a series of *internal SR* methods. With much progress being made, it is recognized that external and internal SR methods each suffer from their certain drawbacks. However, their complementary properties inspire us to propose the *joint super-resolution* (joint SR), that adaptively utilizes both external and internal examples for the SR task. The contributions of this paper are multi-fold:

- We propose joint SR exploiting both external and internal examples, by defining an adaptive combination of different loss functions.

- We apply *epitomic matching* [6] to enforcing self-similarity in SR. Compared the local nearest neighbor (NN) matching adopted in [5], epitomic matching features more robustness to outlier features, as well as the ability to perform efficient non-local searching.
- We carry out a human subjective evaluation survey to evaluate SR result quality based on visual perception, among several state-of-the-art methods.

II. A MOTIVATION STUDY OF JOINT SR

A. Related Work

The joint utilization of both external and internal examples has been most studied for image denoising [17]. Mosseri et. al. [18] first proposed that some image patches inherently prefer internal examples for denoising, whereas other patches inherently prefer external denoising. Such a preference is in essence the tradeoff between noise-fitting versus signal-fitting. Burger et. al. [16] proposed a learning-based approach that automatically combines denoising results from an internal and an external method. The learned combining strategy outperforms both internal and external approaches across a wide range of images, being closer to theoretical bounds.

In SR literature, while the most popular methods are based on either external or internal similarities, there have been limited efforts to utilize one to regularize the other. The authors in [19] incorporated both a local autoregressive (AR) model and a nonlocal self-similarity regularization term, into the sparse representation framework, weighted by constant coefficients. Yang et. al. [20] learned the (approximated) nonlinear SR mapping function from a collection of external images with the help of in-place self-similarity. More recently, an explicitly joint model is put forward in [23], including two loss functions by sparse coding and local scale invariance, bound by an indicator function to decide which loss function will work for each patch of the input image. Despite the existing efforts, there is little understanding on how the external and internal examples interact with each other in SR, how to judge the external versus internal preference for each patch, and how to make them collaborate towards an overall optimized performance.

External SR methods use a universal set of example patches to predict the missing (high-frequency) information for the HR image. In [7], during the training phase, LR-HR patch pairs

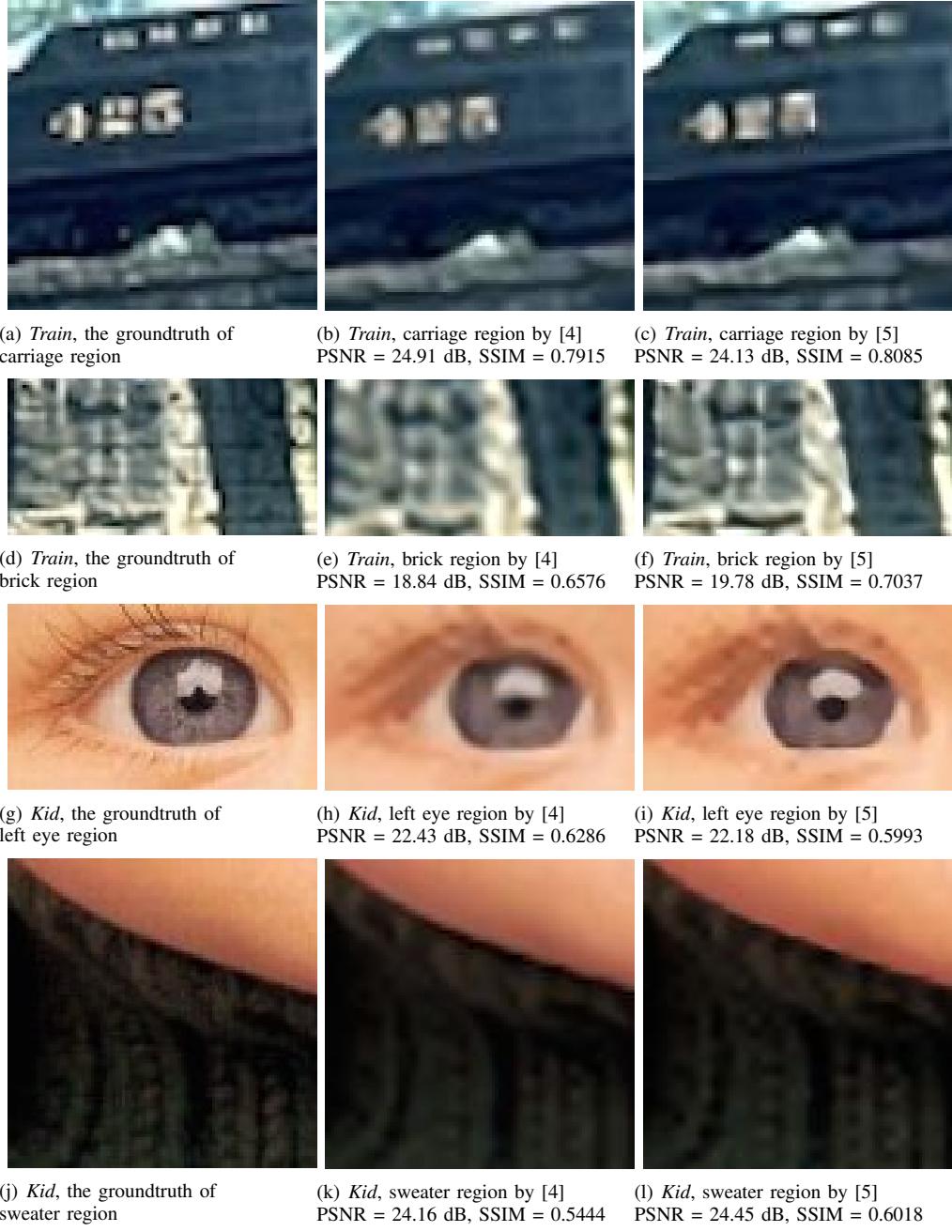


Fig. 1. Visual comparisons of both external and internal SR methods on different image local regions. The PSNR and SSIM values are also calculated and reported.

are collected. Then in the test phase, each input LR patch is found with a nearest neighbor (NN) match in the LR patch pool, and its corresponding HR patch is selected as the output. It is further formulated as a kernel ridge regression (KRR) in [8]. More recently, a popular class of external SR methods are associated with the *sparse coding* technique [9], [10]. The patches of a natural image can be represented as a sparse linear combination of elements within a redundant pre-trained dictionary. Following this principle, the advanced *coupled sparse coding* is further proposed in [4], [10]. External SR methods are known for their capabilities to produce plausible image appearances. However, there is no guarantee that an arbitrary input patch can be well matched or represented

by the external dataset of limited size. When dealing with some unique features that rarely appear in the given dataset, external SR methods are prone to produce either noise or oversmoothness [11]. It constitutes the inherent problem of any external SR method with a finite-size training set [12].

Internal SR methods search for example patches from the input image itself, based on the fact that patches often tend to recur within the image [13], [14], [11], or across different image scales [5]. Although internal examples provide a limited number of references, they are very relevant to the input image. However, this type of approach has a limited performance, especially for irregular patches without any discernible repeating pattern [15]. Due to the insufficient patch pairs, the

mismatches of internal examples often lead to more visual artifacts. In addition, epitome was proposed in [6], [24], [25] to summarize both local and non-local similar patches and reduces the artifacts caused by neighborhood matching. We apply epitome as an internal SR technique in this paper, and evidence its advantages by our experiments.

B. Comparing External and Internal SR Methods

Both external and internal SR methods have different advantages and drawbacks. See Fig. 1 for a few specific examples. The first two rows of images are cropped from the $3 \times$ SR results of the *Train* image, and the last two rows from the $4 \times$ SR results of the *Kid* image. Each row of images are cropped from the same spatial location of the groundtruth image, the SR result by the external method [4], and the SR result by the internal method [5], respectively. In the first row, the top contour of carriage (c) contains noticeable structural deformations, and the numbers “425” are more blurred than those in (b). That is because the numbers can more easily find counterparts or similar structure components from an external dataset; but within the same image, there are few recurring patterns that look visually similar to the numbers. Internal examples generate sharper SR results in images (f) than (e), since the bricks repeat their own patterns frequently, and thus the local neighborhood is rich in internal examples. Another winning case of external examples is between (h) and (i), as in the latter, inconsistent artifacts along the eyelid and around the eyeball are obvious. Because the eye region is composed of complex curves and fine structures, external examples encompass more suitable reference patches and perform a more natural-looking SR. In contrast, the repeating sweater textures lead to a sharper SR in (l) than that in (k). The PSNR and SSIM [26] results are also calculated for all, which further validate our visual observations.

These comparisons display the generally different, even complementary behaviors of external and internal SR. Based on the observations, we expect that the external examples contribute to visually pleasant SR results for smooth regions as well as some irregular structures that barely recur in the input. Meanwhile, internal examples serve as a powerful source to reproduce unique and singular features that rarely appear externally but repeat in the input image (or its different scales). Note that similar arguments have been validated statistically in the the image denoising literature [16].

III. A JOINT SR MODEL

Let \mathbf{X} denote the HR image to be estimated from the LR input \mathbf{Y} . \mathbf{X}_{ij} and \mathbf{Y}_{ij} stand for the (i, j) -th ($i, j = 1, 2, \dots$) patch from \mathbf{X} and \mathbf{Y} , respectively. Considering almost all SR methods work on patches, we define two loss functions $\ell_G(\cdot)$ and $\ell_I(\cdot)$ in a patch-wise manner, which enforce the external and internal similarities, respectively. While one intuitive idea is to minimize a weighted combination of the two loss functions, a patch-wise (adaptive) weight $\omega(\cdot)$ is needed to balance them. We hereby write our proposed joint SR in the

general form:

$$\min_{\mathbf{X}_{ij}, \Theta_G, \Theta_I} \ell_G(\mathbf{X}_{ij}, \Theta_G | \mathbf{Y}_{ij}) + \omega(\Theta_G, \Theta_I | \mathbf{Y}_{ij}) \ell_I(\mathbf{X}_{ij}, \Theta_I | \mathbf{Y}_{ij}). \quad (1)$$

Θ_G and Θ_I are the latent representations of \mathbf{X}_{ij} over the spaces of external and self examples, respectively. The form $f(\mathbf{X}_{ij}, \Theta | \mathbf{Y}_{ij})$, f being ℓ_G , ℓ_I or ω , represents the function dependent on variables \mathbf{X}_{ij} and Θ (Θ_G or Θ_I), with \mathbf{Y}_{ij} known (we omit \mathbf{Y}_{ij} in all formulations hereinafter). We will discuss each component in (1) next.

One specific form of joint SR will be discussed in this paper. However, note that with different choices of $\ell_G(\cdot)$, $\ell_I(\cdot)$, and $\omega(\cdot)$, a variety of methods can be accommodated in the framework. For example, if we set $\ell_G(\cdot)$ as the (adaptively reweighted) sparse coding term, while choosing $\ell_I(\cdot)$ equivalent to the two local and non-local similarity based terms, then (1) becomes the model proposed in [19], with $\omega(\cdot)$ being some empirically chosen constants.

A. Sparse Coding for External Examples

The HR and LR patch spaces $\{\mathbf{X}_{ij}\}$ and $\{\mathbf{Y}_{ij}\}$ are assumed to be tied by some mapping function. With a well-trained coupled dictionary pair $(\mathbf{D}_h, \mathbf{D}_l)$ (see [4] for details on training a coupled dictionary pair), the *coupled sparse coding* [10] assumes that $(\mathbf{X}_{ij}, \mathbf{Y}_{ij})$ tends to admit a common sparse representation \mathbf{a}_{ij} . Since \mathbf{X} is unknown, Yang et. al. [10] suggest to first infer the sparse code \mathbf{a}_{ij}^L of \mathbf{Y}_{ij} with respect to \mathbf{D}_l , and then use it as an approximation of \mathbf{a}_{ij}^H (the sparse code of \mathbf{X}_{ij} with respect to \mathbf{D}_h), to recover $\mathbf{X}_{ij} \approx \mathbf{D}_h \mathbf{a}_{ij}^L$. We set $\Theta_G = \mathbf{a}_{ij}$ and constitute the loss function enforcing external similarity:

$$\ell_G(\mathbf{X}_{ij}, \mathbf{a}_{ij}) = \lambda \|\mathbf{a}_{ij}\|_1 + \|\mathbf{D}_l \mathbf{a}_{ij} - \mathbf{Y}_{ij}\|_F^2 + \|\mathbf{D}_h \mathbf{a}_{ij} - \mathbf{X}_{ij}\|_F^2. \quad (2)$$

B. Epitomic Matching for Internal Examples

1) *The High Frequency Transfer Scheme*: Based on the observation that singular features like edges and corners in small patches tend to repeat almost identically across different image scales, Freedman and Fattal [5] applied the “high frequency transfer” method to searching the high-frequency component for a target HR patch, by NN patch matching across scales. Defining a linear interpolation operator \mathcal{U} and a downsampling operator \mathcal{D} , for the input LR image \mathbf{Y} , we first obtain its initial upsampled image $\mathbf{X}'^E = \mathcal{U}(\mathbf{Y})$, and a smoothed input image $\mathbf{Y}' = \mathcal{D}(\mathcal{U}(\mathbf{Y}))$. Given the smoothed patch \mathbf{X}'^E_{ij} , the missing high-frequency band of each unknown patch \mathbf{X}^E_{ij} is predicted by first solving a NN matching (3):

$$(m, n) = \arg \min_{(m, n) \in \mathcal{W}_{ij}} \|\mathbf{Y}'_{mn} - \mathbf{X}'^E_{ij}\|_F^2, \quad (3)$$

where \mathcal{W}_{ij} is defined as a small local searching window on image \mathbf{Y}' . We could also simply express it as $(m, n) = f_{NN}(\mathbf{X}'^E_{ij}, \mathbf{Y}')$. With the co-located patch \mathbf{Y}_{mn} from \mathbf{Y} , the high-frequency band $\mathbf{Y}_{mn} - \mathbf{Y}'_{mn}$ is pasted onto \mathbf{X}'^E_{ij} , i.e., $\mathbf{X}^E_{ij} = \mathbf{X}'^E_{ij} + \mathbf{Y}_{mn} - \mathbf{Y}'_{mn}$.

2) *EPI: Epitomic Matching for Internal SR*: The matching of $\mathbf{X}_{ij}^{'}E$ over the smoothed input image \mathbf{Y}' makes the core step of the high frequency transfer scheme. However, the performance of NN matching (3) is degraded with the presence of noise and outliers. Moreover, the NN matching in [5] is restricted to a local window for efficiency, which potentially accounts for some rigid artifacts.

Instead, we propose *epitomic matching* to replace NN matching in the above frequency transfer scheme. As a generative model, epitome [25], [27] summarizes a large set of raw image patches into a condensed representation in a way similar to Gaussian Mixture Models. We first learn an epitome $\mathbf{e}_{\mathbf{Y}'}$ from \mathbf{Y}' , and then match each $\mathbf{X}_{ij}^{'}E$ over $\mathbf{e}_{\mathbf{Y}'}$ rather than \mathbf{Y}' directly. Assume $(m, n) = f_{ept}(\mathbf{X}_{ij}^{'}E, \mathbf{e}_{\mathbf{Y}'})$, where f_{ept} denotes the procedure of epitomic matching by $\mathbf{e}_{\mathbf{Y}'}$. It then follows the same way as in [5]: $\mathbf{X}_{ij}^E = \mathbf{X}_{ij}^{'}E + \mathbf{Y}_{mn} - \mathbf{Y}_{mn}^{'}$; the only difference here is the replacement of f_{NN} with f_{ept} . The high-frequency transfer scheme equipped with epitomic matching can thus be applied to SR by itself as well, named *EPI* for short, which will be included in our experiments in Section 4 and compared to the method using NN matching in [5].

Since $\mathbf{e}_{\mathbf{Y}'}$ summarizes the patches of the entire \mathbf{Y}' , the proposed epitomic matching benefits from non-local patch matching. In the absence of self-similar patches in the local neighborhood, epitomic matching weights refer to non-local matches, thereby effectively reducing the artifacts arising from local matching [5] in a restricted small neighborhood. In addition, note that each epitome patch summarizes a batch of similar raw patches in \mathbf{Y}' . For any patch \mathbf{Y}'_{ij} that contains certain noise or outliers in \mathbf{Y}' , its has a small posterior and thus tends not be selected as candidate matches for $\mathbf{X}_{ij}^{'}E$, improving the robustness of matching. The algorithm details of epitomic matching are included in Appendix.

Moreover, we can also incorporate Nearest Neighbor (NN) matching to our epitomic matching, leading to a enhanced patch matching scheme that features both non-local (by epitome) and local (by NN) matching. Suppose the high frequency components obtained by epitomic matching and NN matching for patch $\mathbf{X}_{ij}^{'}E$ are $\mathbf{H}_{ij,e}$ and $\mathbf{H}_{ij,NN}$ respectively, we use a smart weighted average of the two as the final high frequency component \mathbf{H}_{ij} :

$$\mathbf{H}_{ij} = w\mathbf{H}_{ij,e} + (1-w)\mathbf{H}_{ij,NN} \quad (4)$$

where the weight $w = p(\mathcal{T}_{ij}^* | \mathbf{X}_{ij}^{'}E, \mathbf{e})$ denotes the probability of the most probable hidden mapping given the patch $\mathbf{X}_{ij}^{'}E$. A higher w indicates that the patch $\mathbf{X}_{ij}^{'}E$ is more likely to have a reliable match by epitomic matching (with the probability measured through the corresponding most probable hidden mapping), thereby a larger weight is associated with the epitomic matching, and vice versa. This is the practical implementation of EPI that we used in the paper.

Finally, we let $\Theta_I = \mathbf{X}_{ij}^E$ and define

$$\ell_{\mathcal{I}}(\mathbf{X}_{ij}, \mathbf{X}_{ij}^E) = \|\mathbf{X}_{ij} - \mathbf{X}_{ij}^E\|_F^2, \quad (5)$$

where \mathbf{X}_{ij}^E is the internal SR result by epitomic matching.

C. Learning the Adaptive Weights

In [18], Mosseli et.al. showed that the internal versus external preference is tightly related to the Signal-to-Noise-Ratio (SNR) estimate of each patch. Inspired by that finding, we could seek similar definitions of "noise" in SR based on the latent representation errors. The *external noise* is defined by the residual of sparse coding

$$N_g(\mathbf{a}_{ij}) = \|\mathbf{D}_1\mathbf{a}_{ij} - \mathbf{Y}_{ij}\|_F^2. \quad (6)$$

Meanwhile, the *internal noise* finds its counterpart definition by the epitomic matching error within f_{ept} :

$$N_i(\mathbf{X}_{ij}^E) = \|\mathbf{Y}'_{mn} - \mathbf{X}_{ij}^{'}E\|_F^2, \quad (7)$$

where \mathbf{Y}'_{mn} is the matching patch in \mathbf{Y}' for $\mathbf{X}_{ij}^{'}E$.

Usually, the two "noises" are on the same magnitude level, which aligns with the fact that external- and internal-examples will have similar performances on many (such as homogenous regions). However, there do exist patches where the two have significant differences in performances, as shown in Fig. 1, which means the patch has a strong preference toward one of them. In such cases, the "preferred" term needs to be sufficiently emphasized. We thus construct the following patch-wise adaptive weight (p is the hyperparameter):

$$\omega(\alpha_{ij}, \mathbf{X}_{ij}^E) = \exp(p \cdot [N_g(\mathbf{a}_{ij}) - N_i(\mathbf{X}_{ij}^E)]). \quad (8)$$

When the internal noise becomes larger, the weight decays quickly to ensure that external similarity dominates, and vice versa.

D. Optimization

Directly solving (1) is very complex due to the its high nonlinearity and entanglement among all variables. Instead, we follow the coordinate descent fashion [28] and solve the following three sub-problems iteratively.

1) **\mathbf{a}_{ij} -subproblem:** Fixing \mathbf{X}_{ij} and \mathbf{X}_{ij}^E , we have the following minimization w.r.t α_{ij}

$$\begin{aligned} \min_{\mathbf{a}_{ij}} \quad & \lambda \|\mathbf{a}_{ij}\|_1 + \|\mathbf{D}_1\mathbf{a}_{ij} - \mathbf{Y}_{ij}\|_F^2 + \|\mathbf{D}_h\mathbf{a}_{ij} - \mathbf{X}_{ij}\|_F^2 \\ & + [\ell_{\mathcal{I}}(\mathbf{X}_{ij}, \mathbf{X}_{ij}^E) \cdot \exp(-p \cdot N_i(\mathbf{X}_{ij}^E))] \cdot \exp(p \cdot N_g(\mathbf{a}_{ij})). \end{aligned} \quad (9)$$

The major bottleneck of exactly solving (9) lies in the last exponential term. We let \mathbf{a}_{ij}^0 denote the \mathbf{a}_{ij} value solved in the last iteration. We then apply first-order Taylor expansion to the last term of the objective in (9), with regard to $N_g(\mathbf{a}_{ij})$ at $\alpha_{ij} = \mathbf{a}_{ij}^0$, and solve the approximated problem as follows:

$$\min_{\mathbf{a}_{ij}} \quad \lambda \|\mathbf{a}_{ij}\|_1 + (1+C) \|\mathbf{D}_1\mathbf{a}_{ij} - \mathbf{Y}_{ij}\|_F^2 + \|\mathbf{D}_h\mathbf{a}_{ij} - \mathbf{X}_{ij}\|_F^2, \quad (10)$$

where C is the constant coefficient:

$$\begin{aligned} C &= [\ell_{\mathcal{I}}(\mathbf{X}_{ij}, \mathbf{X}_{ij}^E) \cdot \exp(-p \cdot N_i(\mathbf{X}_{ij}^E))] \cdot [p \cdot \exp(p \cdot N_g(\mathbf{a}_{ij}^0))] \\ &= p \ell_{\mathcal{I}}(\mathbf{X}_{ij}, \mathbf{X}_{ij}^E) \cdot \omega(\alpha_{ij}^0, \mathbf{X}_{ij}^E). \end{aligned} \quad (11)$$

(10) can be conveniently solved by the feature sign algorithm [9]. Note (10) is a valid approximation of (9) since \mathbf{a}_{ij} and \mathbf{a}_{ij}^0 become quite close after a few iterations, so that the higher-order Taylor expansions can be reasonably ignored.

Another noticeable fact is that since $C > 0$, the second term is always emphasized more than the third term, which makes sense as \mathbf{Y}_{ij} is the “accurate” LR image, while \mathbf{X}_{ij} is just an estimate of the HR image and is thus less weighted. Further considering the formulation (11), C grows up as $\omega(\alpha_{ij}^0, \mathbf{X}_{ij}^E)$ turns larger. That implies when external SR becomes the major source of “SR noise” on a patch in the last iteration, (10) will accordingly rely less on the last solved \mathbf{X}_{ij} .

2) \mathbf{X}_{ij}^E -subproblem: Fixing \mathbf{a}_{ij} and \mathbf{X}_{ij} , the \mathbf{X}_{ij}^E -subproblem becomes

$$\min_{\mathbf{X}_{ij}^E} \exp(-p \cdot \|\mathbf{Y}'_{mn} - \mathbf{X}_{ij}^{E'}\|_F^2) \ell_{\mathcal{I}}(\mathbf{X}_{ij}, \mathbf{X}_{ij}^E), \quad (12)$$

While in Section III.B.2, $X_{i,j}^E$ is directly computed from the input LR image, the objective in (12) is dependent on not only $X_{i,j}^E$ but also $X_{i,j}$, which is not necessarily minimized by the best match $X_{i,j}^E$ obtained from solving f_{ept} . In our implementation, the K best candidates ($K = 5$) that yield minimum matching errors of solving f_{ept} are first obtained. Among all those candidates, we further select the one that minimizes the loss value as defined in (12). By this discrete search-type algorithm, $X_{i,j}^E$ becomes a latent variable to be updated together with $X_{i,j}$ per iteration, and is better suited for the global optimization than the simplistic solution by solving f_{ept} .

3) \mathbf{X}_{ij} -subproblem: With both \mathbf{a}_{ij} and \mathbf{X}_{ij}^E fixed, the solution of \mathbf{X}_{ij} simply follows a weight least square (WLS) problem:

$$\min_{\mathbf{X}_{ij}} \|\mathbf{D}_h \mathbf{a}_{ij} - \mathbf{X}_{ij}\|_F^2 + \omega(\mathbf{a}_{ij}, \mathbf{X}_{ij}^E) \|\mathbf{X} - \mathbf{X}_{ij}^E\|_F^2, \quad (13)$$

with an explicit solution:

$$\mathbf{X}_{ij} = \frac{\mathbf{D}_h \mathbf{a}_{ij} + \omega(\alpha_{ij}, \mathbf{X}_{ij}^E) \cdot \mathbf{X}_{ij}^E}{1 + \omega(\mathbf{a}_{ij}, \mathbf{X}_{ij}^E)}. \quad (14)$$

IV. EXPERIMENTS

A. Implementation Details

We itemize the parameter and implementation settings for the following group of experiments:

- We use 5×5 patches with one pixel overlapping for all experiments except those on SHD images in Section 4.4, where the patch size is 25×25 with five pixel overlapping.
- In (2), we adopt the \mathbf{D}_l and \mathbf{D}_h trained in the same way as in [4], due to the similar roles played by the dictionaries in their formulation and our ℓ_G function. However, we are aware that such D_l and D_h are not optimized for the proposed method, and will integrate a specifically designed dictionary learning part in future work. λ is empirically set as 1.
- In (5), the size of the epitome is $\frac{1}{4}$ of the image size.
- In (11), we set $p = 1$ for all experiments. We also observed in experiments that a larger p will usually lead to a faster decrease in objective value, but the SR result quality may degrade a bit.
- We initialize \mathbf{a}_{ij} by solving coupled sparse coding in [4]. \mathbf{X}_{ij} is initialized by bicubic interpolation.

- We set the maximum iteration number to be 10 for the coordinate descent algorithm. For SHD cases, the maximum iteration number is adjusted to be 5.
- For color images, we apply SR algorithms to the illuminance channel only, as humans are more sensitive to illuminance changes. We then interpolate the color layers (Cb, Cr) using plain bi-cubic interpolation.

B. Comparison with State-of-the-Art Results

We compare the proposed method with the following selection of competitive methods as follows,

- *Bi-Cubic Interpolation (“BCI” for short and similarly hereinafter)*, as a comparison baseline.
- *Coupled Sparse Coding (CSC)* [4], as the classical external-example-based SR.
- *Local Self-Example based SR (LSE)* [5], as the classical internal-example-based SR.
- *Epitome-based SR (EPI)*. We compare EPI to LSE to demonstrate the advantage of epitomic matching over the local NN matching.
- *SR based on In-place Example Regression (IER)* [20], as the previous SR utilizing both external and internal information.
- *The proposed joint SR (JSR)*.

We list the SR results (best viewed on a high-resolution display) for two test images: *Temple* and *Train*, by an amplifying factor of 3. PSNR and SSIM measurements, as well as zoomed local regions (using nearest neighbor interpolation), are available for different methods as well.

In Fig. 2, although greatly outperforming the naive BCI, the external-example based CSC tends to lose many fine details. In contrast, LSE brings out an overly sharp SR result with observable blockiness. EPI produces a more visually pleasing result, through searching for the matches over the entire input efficiently by the pre-trained epitome rather than a local neighborhood. Therefore, EPI substantially reduces the artifacts compared to LSE. But without any external information available, it is still incapable of inferring enough high-frequency details from the input solely, especially under a large amplifying factor. The result of IER greatly improves but is still accompanied with occasional small artifacts. Finally, JSR provides a clear recovery of the steps, and it reconstructs the most pillar textures. In Fig. 3, JSR is the only algorithm which clearly recovers the number on the carrier and the bricks on the bridge simultaneously. The performance superiorities of JSR are also verified by the PSNR comparisons, where larger margins are obtained by JSR over others in both cases.

Next, we move on to the more challenging $4 \times$ SR case, using the *Chip* image which is quite abundant in edges and textures. Since we have no ground truth for the *Chip* image of $4 \times$ size, only visual comparisons are presented. Given such a large SR factor, the CSC result is a bit blurry around the characters on the surface of chip. Both LSE and EPI create jaggy artifacts along the long edge of the chip, as well as small structure distortions. The IER result causes less artifacts but in sacrifice of detail sharpness. The JSR result presents the best SR with few artifacts.

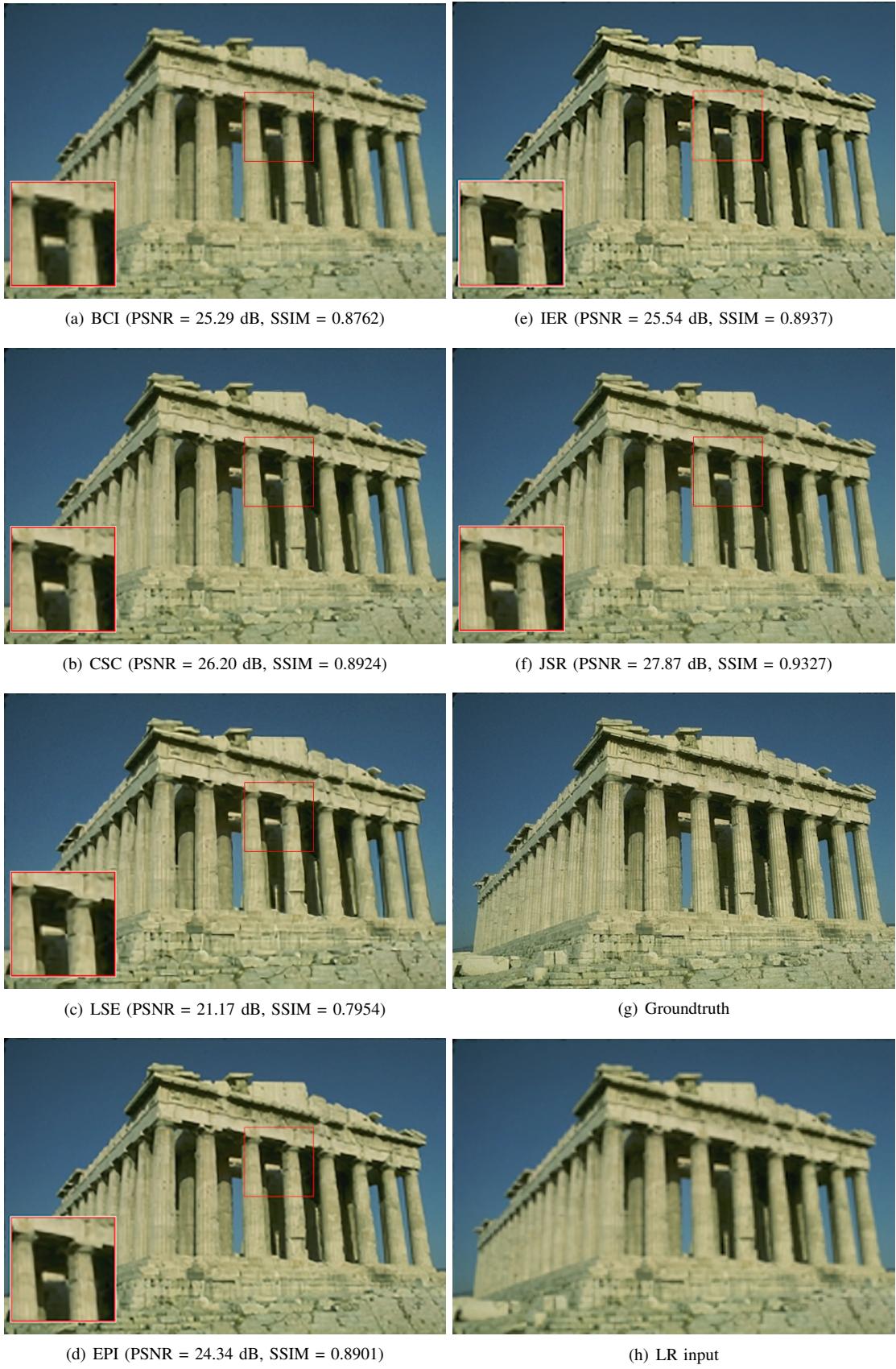


Fig. 2. 3 \times SR results of the *Temple* image.



Fig. 3. 3 \times SR results of the *Train* image.

TABLE II
THE PSNR VALUES (DB) WITH VARIOUS FIXED GLOBAL WEIGHTS (PSNR = 24.1734 DB WITH AN ADAPTIVE WEIGHT)

$\omega = 0.1$	$\omega = 1$	$\omega = 3$	$\omega = 5$	$\omega = 10$
23.13	23.23	23.32	22.66	21.22

The key idea of JSR is utilizing the complementary behavior of both external and internal SR methods. Note when one inverse problem is better solved, it also makes a better parameter estimate for solving the other. JSR is not a simple static weighted average of external SR (CSC) and internal SR (EPI). When optimized jointly, the external and internal subproblems can "boost" each other (through auxiliary variables), and each performs better than being applied independently. That is why JSR gets details that exist in neither internal or external SR result.

To further verify the superiority of JSR numerically, we compare the average PSNR and SSIM results of a few recently-proposed, state-of-the-art single image SR methods, including CSC, LSE, the Adjusted Anchored Neighborhood Regression (A+) [22], and the latest Super-Resolution Convolutional Neural Network (SRCNN) [21]. Table I reports the results on the widely-adopted Set 5 and Set 14 datasets, in terms of both PSNR and SSIM. First, it is not a surprise to us, that JSR does not always yield higher PSNR than SRCNN, et. al., as the epitomic matching component is not meant to be optimized under Mean-Square-Error (MSE) measure, in contrast to the end-to-end MSE-driven regression adopted in SRCNN. However, it is notable that JSR is particularly more favorable by SSIM than other methods, owing to the self-similar examples that convey input-specific structural details. Considering SSIM measures image quality more consistently with human perception, the observation is in accordance with our human subject evaluation results (see Section IV. E).

C. Effect of Adaptive Weight

To demonstrate how the proposed joint SR will benefit from the learned adaptive weight (11), we compare $4 \times$ SR results of *Kid* image, between joint SR solving (1), and its counterpart with fixed global weights , i.e. set the weight ω as constant for all patches. Table 1 shows that the joint SR with an adaptive weight gains a consistent PSNR advantage over the SR with a large range of fixed weights.

More interestingly, we visualize the patch-wise weight maps of joint SR results in Fig. 2 - 4, using heat maps, as in Fig. 5. The (i, j) -th pixel in the weight map denote the final weight of \mathbf{X}_{ij} when the joint SR reaches a stable solution. All weights are normalized between [0,1], by the form of sigmoid function: $\frac{1}{1+\omega(\alpha_{ij}, \mathbf{X}_{ij}^E)}$, for visualization purpose. A larger pixel value in the weight maps denote a smaller weight and thus a higher emphasis on external examples, and vice versa. For *Temple* image, Fig. 5 (a) clearly manifests that self examples dominate the SR of the temple building that is full of texture patterns. Most regions of Fig. 5 (b) are close to 0.5, which means that $\omega(\alpha_{ij}, \mathbf{X}_{ij}^E)$ is close to 1 and external and internal examples have similar performances on

most patches. However, internal similarity makes more significant contributions in reconstructing the brick regions, while external examples works remarkably better on the irregular contours of forests. Finally, the *Chip* image is an example where external examples have advantages on the majority of patches. Considering self examples prove to create artifacts here (see Fig. 4 (c) (d)), they are avoided in joint SR by the adaptive weights.

D. SR Beyond Standard Definition: From HD Image to UHD Image

In almost all SR literature, experiments are conducted with Standard-Definition (SD) images (720×480 or 720×576 pixels) or smaller. The High-Definition (HD) formats: 720p (1280×720 pixels) and 1080p (1920×1080 pixels) have become popular today. Moreover, Ultra High-Definition (UHD) TVs are hitting the consumer markets right now with the 3840×2160 resolution. It is thus quite interesting to explore whether SR algorithms established on SD images can be applied or adjusted for HD or UHD cases. In this section, we upscale HD images of 1280×720 pixels to UHD results of 3840×2160 pixels, using competitor methods and our joint SR algorithm.

Since most HD and UHD images typically contain much more diverse textures and a richer collection of fine structures than SD images, we enlarge the patch size from 5×5 to 25×25 (the dictionary pair is therefore re-trained as well) to capture more variations, meanwhile increasing the overlapping from one pixel to five pixels to ensure enough spatial consistency. Hereby JSR is compared with its two "component" algorithms, i.e., CSC and EPI. We choose several challenging SHD images (3840×2160 pixels) with very cluttered texture regions, downsampling them to HD size (1280×720 pixel) on which we apply the SR algorithm with a factor of 3. In all cases, our results are consistently sharper and clearer. The SR results (zoomed local regions) of the *Leopard* image are displayed in Fig. 6 for examples, with the PSNR and SSIM measurements of full-size results.

E. Subjective Evaluation

We conduct an online *subjective evaluation* survey ¹ on the quality of SR results produced by all different methods in Section 4.2. Ground truth HR images are also included when they are available as references. Each participant of the survey is shown a set of HR image pairs obtained using two different methods for the same LR image. For each pair, the participant needs to decide which one is better than the other in terms of perceptual quality. The image pairs are drawn from all the competitive methods randomly, and the images winning the pairwise comparison will be compared again in the next round, until the best one is selected. We have a total of 101 participants giving 1,047 pairwise comparisons, over six images which are commonly used as benchmark images in SR, with different scaling factors (*Kid* $\times 4$, *Chip* $\times 4$, *Statue* $\times 4$, *Leopard* $\times 3$, *Temple* $\times 3$ and *Train* $\times 3$). We fit a Bradley-Terry

¹<http://www.ifp.illinois.edu/~wang308/survey>

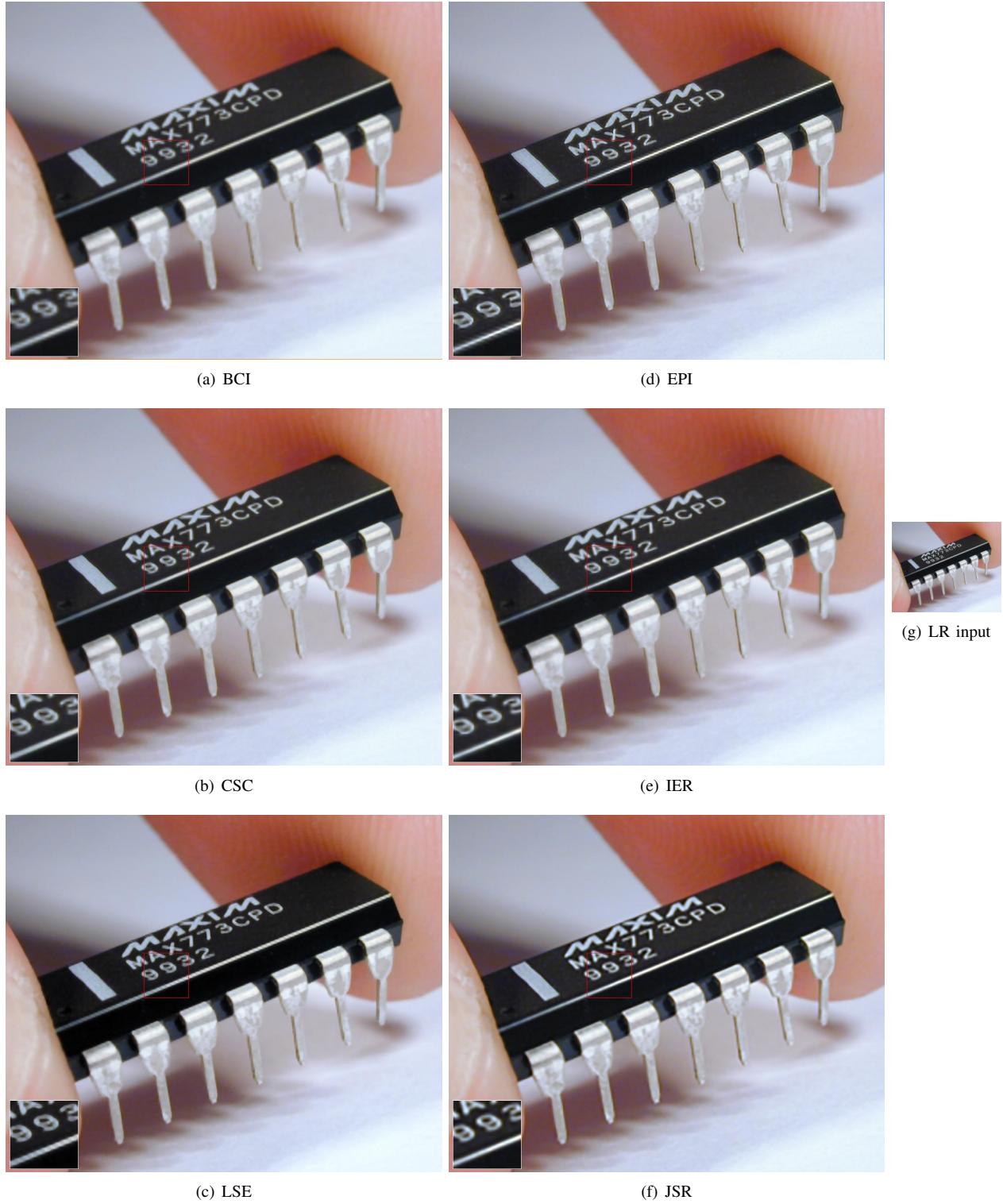


Fig. 4. $4 \times$ SR results of the *Chip* image.

TABLE I
AVERAGE PSNR (DB) AND SSIM PERFORMANCES COMPARISONS ON THE SET 5 AND SET 14 DATASETS

		Bicubic	Sparse Coding [10]	Freedman et.al. [5]	A+ [22]	SRCCN [21]	JSR
<i>Set 5, st=2</i>	PSNR	33.66	35.27	33.61	36.24	36.66	36.71
	SSIM	0.9299	0.9540	0.9375	0.9544	0.9542	0.9573
<i>Set 5, st=3</i>	PSNR	30.39	31.42	30.77	32.59	32.75	32.54
	SSIM	0.8682	0.8821	0.8774	0.9088	0.9090	0.9186
<i>Set 14, st=2</i>	PSNR	30.23	31.34	31.99	32.58	32.45	32.54
	SSIM	0.8687	0.8928	0.8921	0.9056	0.9067	0.9082
<i>Set 14, st=3</i>	PSNR	27.54	28.31	28.26	29.13	29.60	29.49
	SSIM	0.7736	0.7954	0.8043	0.8188	0.8215	0.8242

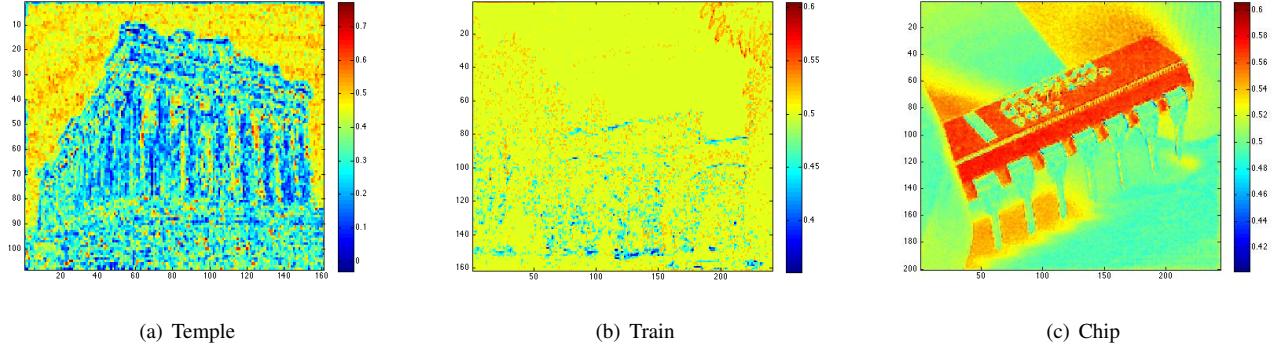


Fig. 5. The weight maps of (a) *Temple* image; (b) *Train* image; (c) *Chip* image.

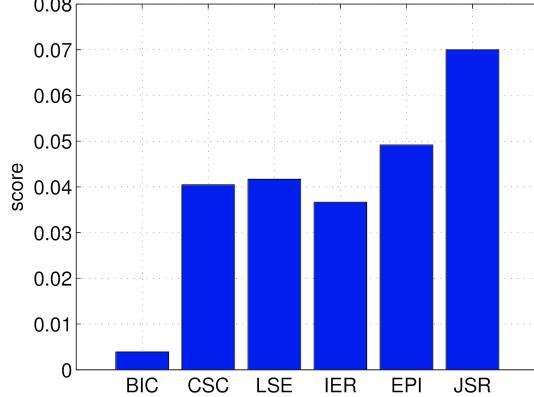


Fig. 7. Subjective SR quality scores for different methods. The ground truth has score 1.

[29] model to estimate the subjective scores for each method so that they can be ranked. More experiment details are included in our Appendix. Figure 7 shows the estimated scores for the six SR methods in our evaluation. As expected, all SR methods receive much lower scores compared to ground truth (set as score 1), showing the huge challenge of the SR problem itself. Also, the bicubic interpolation is significantly worse than others. The proposed JSR method outperforms all other state-of-the-art methods by a large margin, which proves that JSR can produce more visually favorable HR images by human perception.

V. CONCLUSION

This paper presents a joint single image SR model, by learning from both external and internal examples. We define the two loss functions by sparse coding and epitomic matching, respectively, and construct an adaptive weight to balance the

two terms. Experimental results demonstrate that joint SR outperforms existing state-of-the-art methods for various test images of different definitions and scaling factors, and is also significantly more favored by user perception. We will further integrate dictionary learning into the proposed scheme, as well as reducing its complexity.

APPENDIX

1. Epitomic Matching Algorithm

We assume an epitome \mathbf{e} of size $M_e \times N_e$, for an input image of size $M \times N$, where $M_e < M$ and $N_e < N$. Similarly to GMMs, \mathbf{e} contains three parameters [6], [24], [25]: μ , the Gaussian mean of size $M_e \times N_e$; ϕ , the Gaussian variance of size $M_e \times N_e$; and π , the mixture coefficients. Suppose Q densely sampled, overlapped patches from the input image, i.e. $\{\mathbf{Z}_k\}_{k=1}^Q$. Each \mathbf{Z}_k contains pixels with image coordinates \mathbf{S}_k , and is associated with a hidden mapping \mathcal{T}_k from \mathbf{S}_k to the epitome coordinates. All the Q patches are generated independently from the epitome and the corresponding hidden mappings as below:

$$\prod_{k=1}^Q p(\{\mathbf{Z}_k\}_{k=1}^Q | \{\mathcal{T}_k\}_{k=1}^Q, \mathbf{e}) = \prod_{k=1}^Q p(\mathbf{Z}_k | \mathcal{T}_k, \mathbf{e}) \quad (15)$$

The probability $p(\mathbf{Z}_k | \mathcal{T}_k, \mathbf{e})$ in (15) is computed by the Gaussian distribution where the Gaussian component is specified by the hidden mapping \mathcal{T}_k . \mathcal{T}_k behaves similar to the hidden variable in the traditional GMMs.

Figure 8 illustrates the role that the hidden mapping plays in the epitome as well as the graphical model illustration for epitome. With all the above notations, our goal is to find the epitome \mathbf{e} that maximizes the log likelihood function



(a) Full SHD image

(b) Local region from SR result by BCI
PSNR = 24.14 dB, SSIM = 0.9701(d) Local region from SR result by EPI
PSNR = 23.58 dB, SSIM = 0.9656(c) Local region from SR result by CSC
PSNR = 25.32 dB, SSIM = 0.9618(e) Local region from SR result by JSR
PSNR = 25.82 dB, SSIM = 0.9746Fig. 6. $3 \times$ SR results of the *Leopard* image (local region displayed).

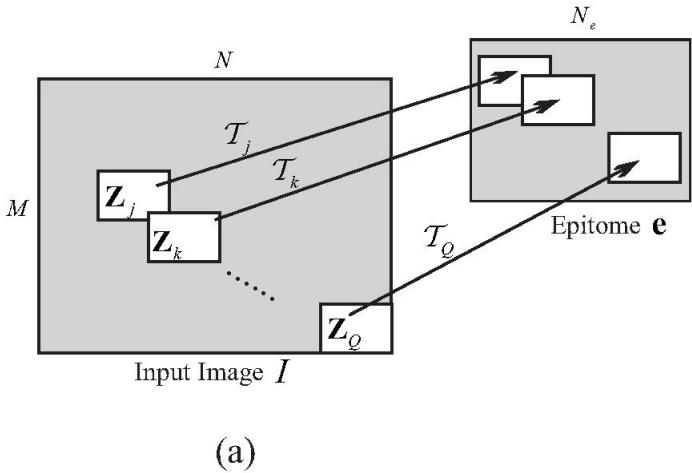


Fig. 8. (a) The hidden mapping T_k maps the image patch Z_k to its corresponding patch of the same size in e , and Z_k can be mapped to any possible epitome patch in accordance with T_k . (b) The epitome graphical model

$\mathbf{e} = \arg \max_{\hat{\mathbf{e}}} \log p(\{\mathbf{Z}_k\}_{k=1}^Q | \hat{\mathbf{e}})$, which can be solved by the Expectation-Maximization (EM) algorithm [6], [27].

The Expectation step in the EM algorithm which computes the posterior of all the hidden mappings accounts for the most time consuming part of the learning process. Since the posterior of the hidden mappings for all the patches are independent of each other, they can be computed in parallel. Therefore, the learning process can be significantly accelerated by parallel computing.

With the epitome $\mathbf{e}_{Y'}$ learned from the smoothed input image \mathbf{Y}' , the location of the matching patch in the epitome $\mathbf{e}_{Y'}$ for each patch \mathbf{X}'_{ij}^E is specified by the most probable hidden mapping for \mathbf{X}'_{ij}^E :

$$\mathcal{T}_{ij}^* = \arg \max_{\mathcal{T}_{ij}} p(\mathcal{T}_{ij} | \mathbf{X}'_{ij}^E, \mathbf{e}) \quad (16)$$

The top K patches in \mathbf{Y}' with large posterior probabilities $p(\mathcal{T}_{ij}^* | \cdot, \mathbf{e})$ are regarded as the candidate matches for the patch \mathbf{X}'_{ij} , and the match \mathbf{Y}'_{mn} is the one in these K candidate patches which has minimum Sum of Squared Distance (SSD) to \mathbf{X}'_{ij}^E . Note that the indices of the K candidate patches in \mathbf{Y}' for each epitome patch are pre-computed and stored when training the epitome $\mathbf{e}_{Y'}$ from the smoothed input image \mathbf{Y}' , which makes epitomic matching efficient.

EPI significantly reduces the artifacts and produces more visually pleasing SR results by the dynamic weighting (4), compared to the local NN matching method [5].

2. Subjective Review Experiment

The methods under comparison include BIC, CSC, LSE, IER, EPI, JSR. Ground truth HR images are also included when they are available as references. Each of the human subject participating in the evaluation is shown a set of HR image pairs obtained using two different methods for the same LR image. For each pair, the subject needs to decide which one is better than the other in terms of perceptual quality. The image pairs are drawn from all the competitive methods randomly, and the images winning the pairwise comparison

will be compared again in the next round until the best one is selected.

We have a total of 101 participants giving 1,047 pairwise comparisons over 6 images with different scaling factors (“Kid” $\times 4$, “Chip” $\times 4$, “Statue” $\times 4$, “Leopard” $\times 3$, “Temple” $\times 3$ and “Train” $\times 3$). Not every participant completed all the comparisons but their partial responses are still useful. All the evaluation results can be summarized into a 7×7 winning matrix \mathbf{W} for 7 methods (including ground truth), based on which we fit a Bradley-Terry [29] model to estimate the subjective score for each method so that they can be ranked. In the Bradley-Terry model, the probability that an object X is favored over Y is assumed to be

$$p(X \succ Y) = \frac{e^{s_X}}{e^{s_X} + e^{s_Y}} = \frac{1}{1 + e^{s_Y - s_X}}, \quad (17)$$

where s_X and s_Y are the subjective scores for X and Y . The scores s for all the objects can be jointly estimated by maximizing the log likelihood of the pairwise comparison observations:

$$\max_{\mathbf{s}} \sum_{i,j} w_{ij} \log \left(\frac{1}{1 + e^{s_j - s_i}} \right), \quad (18)$$

where w_{ij} is the (i, j) -th element in the winning matrix \mathbf{W} , representing the number of times when method i is favored over method j . We use the Newton-Raphson method to solve Eq. (18) and set the score for ground truth method as 1 to avoid the scale issue.

Fig. 7 shows the estimated scores for six SR methods in our evaluation. As expected, all the SR methods have much lower scores than ground truth, showing the great challenge in SR problem. Also, the bicubic interpolation is significantly worse than other SR methods. The proposed JSR method outperforms other previous state-of-the-art methods by a large margin, which verifies that JSR can produce visually more pleasant HR images than other approaches.

REFERENCES

- [1] S. C. Park, M. K. Park, and K. G. Kang, “Super-resolution image reconstruction: A technical overview,” *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21-36, 2003.
- [2] R. Fattal, “Image upsampling via imposed edge statistics,” *ACM Transactions on Graphics*, vol. 26, no. 3, pp. 95-102, 2007.
- [3] Z. Lin, and H. Y. Shum, “Fundamental limits of reconstruction-based superresolution algorithms under local translation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 83-97, 2004.
- [4] J. Yang, Z. Wang, Z. Lin, S. Chen, and T. Huang, “Coupled Dictionary Training for Image Super-resolution,” *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3467-3478, 2012.
- [5] G. Freedman and R. Fattal, “Image and video upscaling from local self-examples,” *ACM Transactions on Graphics*, vol. 28, no. 3, 2010.
- [6] N. Jojic, B. J. Frey, and A. Kannan, “Epitomic analysis of appearance and shape,” in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 34-43, 2003.
- [7] W. T. Freeman, T. R. Jones, and E. C. Pasztor, “Example-based super-resolution,” *IEEE Computer Graphics and Applications*, vol. 22, no. 2, pp. 56-65, 2002.
- [8] K. I. Kim, and Y. Kwon, “Single-image super-resolution using sparse regression and natural image prior,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 1127-1133, 2010.
- [9] H. Lee, A. Battle, R. Raina, and A. Y. Ng, “Efficient sparse coding algorithms,” in *Proceedings of Neural Information Processing Systems (NIPS)*, pp. 801-808, 2007.

- [10] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861-2873, 2010.
- [11] C. Yang, J. Huang, and M. Yang, "Exploiting self-similarities for single frame super-resolution," in *Proceedings of Asian Conference on Computer Vision (ACCV)*, pp. 1807-1818, 2010.
- [12] W. Dong, L. Zhang, G. Shi and X. Li, "Nonlocally centralized sparse representation for image restoration," *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1620-1630, 2012.
- [13] D. Glasner, S. Bagon and M. Irani, "Super-resolution from a single image," in *Proceeding of IEEE International Conference on Computer Vision (ICCV)*, pp. 349-356, 2009.
- [14] J. Mairal, F. Bach, J. Ponce, G. Sapiro and A. Zisserman, "Non-local sparse models for image restoration," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 2272-2279, 2009.
- [15] P. Chatterjee, and P. Milanfar, "Is denoising dead?" *IEEE Transactions on Image Processing*, vol. 19, no. 4, pp. 895-911, 2010.
- [16] H. Burger, C. Schuler, and S. Harmeling, "Learning how to combine internal and external denoising methods," *Lecture Notes in Computer Science: Pattern Recognition*, vol. 8142, pp. 121-130, 2013.
- [17] M. Zontak, "Internal statistics of a single natural image," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 977-984, 2011.
- [18] I. Mosserei, M. Zontak, and M. Irani, "Combining the power of internal and external denoising", *IEEE International Conference on Computational Photography (ICCP)*, pp. 1-9, 2013.
- [19] W. Dong, D. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization", *IEEE Transactions on Image Processing*, vol. 20, no. 7, pp. 1838-1857, 2011.
- [20] J. Yang, Z. Lin, and S. Cohen, "Fast image super-resolution based on in-place example regression," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1059-1066, 2013.
- [21] C. Dong, C. C. Loy, K. He and X. Tang, "Learning a deep convolutional network for image super-resolution", in *Proceedings of European Conference on Computer Vision (ECCV)*, pp.184-199, 2014.
- [22] R. Tomofte, V. De Smet and L. Van Gool, "A+: Adjusted Anchored Neighborhood Regression for Fast Super-Resolution".
- [23] Z. Wang, Z. Wang, S. Chang, J. Yang and T. S. Huang, "A Joint Perspective Towards Image Super-Resolution: Unifying External- and Self-Examples," accepted by IEEE Winter conference on Applications of Computer Vision(WACV), 2014.
- [24] K. Ni, A. Kannan, A. Criminisi, and J. Winn, "Epitomic location recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2158-2167, 2009.
- [25] X. Chu, S. Yan, L. Li, K. Chan, and T. Huang, "Spatialized epitome and its applications", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 311-318, 2010.
- [26] Wang, Zhou, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. "Image quality assessment: from error visibility to structural similarity." *Image Processing, IEEE Transactions on* 13, no. 4 (2004): 600-612.
- [27] Y. Yang, X. Chu, T. Ng, A. Chia, J. Yang, H. Jin and T. Huang, "Epitomic image colorization", in *Proceedings of IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.
- [28] D. Bertsekas, *Nonlinear Programming*, 2nd ed. Nashua, NH: Athena Scientific, 1999.
- [29] R. A. Bradley, and M. E. Terry, "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika*, pp. 324-345, 1952.